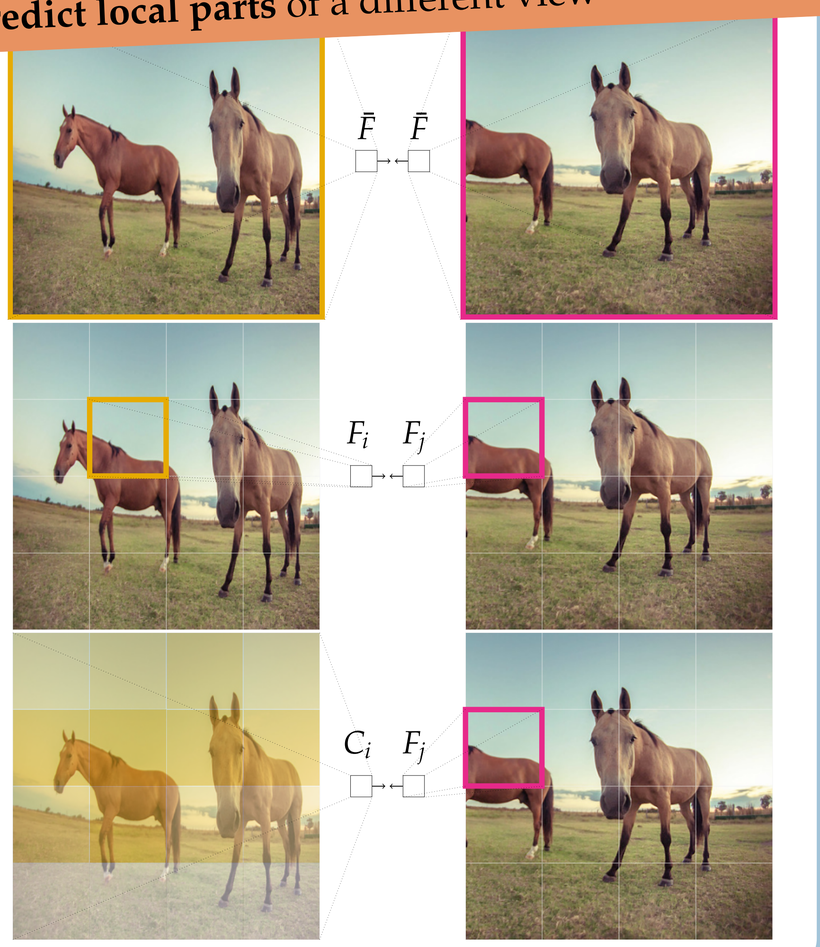
Self-supervised Learning of Contextualized Local Visual Embeddings

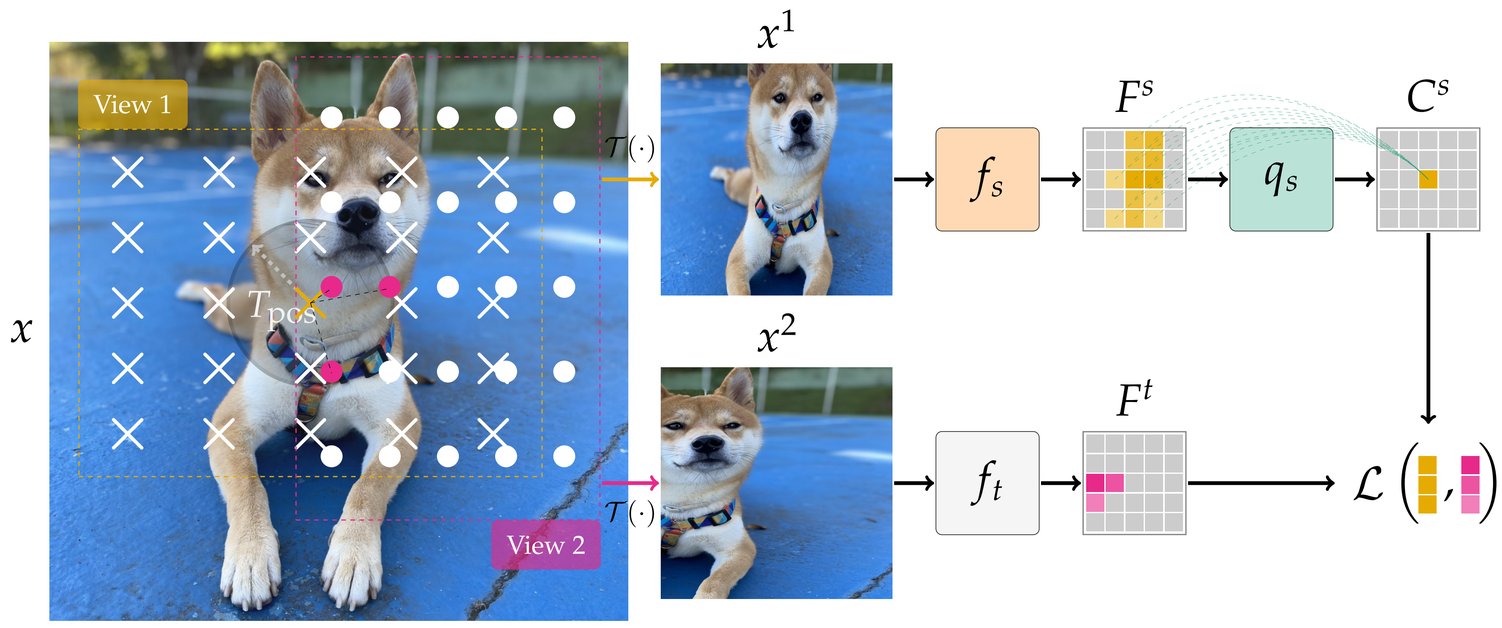
Thalles Silva ❁ Helio Pedrini ❁ Adín Ramírez Rivera ❄

❁ University of Campinas ❄ University of Oslo

Goal Learn representations that **preserve local information** from the input by finding correlations among similar regions of a view to **predict local parts** of a different view



Methodology



- We create views from unlabeled images using random transformations such as flip, color distortions, and cropping.
- Each view is forwarded through a student encoder f_s and a teacher encoder f_t.
- Encoders are composed of a feature extractor, e.g., a CNN, and a MLP projection head.
- For each view, we obtain a tensor of projected local feature maps F taken from the last layers of the CNN encoder (before average pooling).

Obtaining self-supervised dense targets

Strategy match local features based on the pixel's spatial localities

- I^1 and I^2 are lists of 2D points in the pixel space for each view
- For each point I_i^1, we look for pixel correspondences in I^2 to create M
- M is the set of pairs (I_i^1, I_j^2) such that the euclidean distance between points I_i^1 and I_j^2 is smaller than a threshold T_{pos}

$$M = \left\{ \left(I_i^1, I_j^2 \right) : d\left(I_i^1, I_j^2 \right) < T_{\text{pos}} \right\}, \tag{1}$$

- We map points in M from the pixel space to the feature space, to obtain a pair of indices representing matching features from the two views.

Learning long-range dependencies

Combine highly similar local features into a single **contextualized vector** to predict the local target embedding from the second view

- Our proposed predictor head $q_s(F^s) = \text{NMHSA}(F^s)$ receives the projected feature maps F from the student and applies a Normalized Multi-Head Self-Attention (NMHSA) layer to obtain $C^s = q_s(F^s)$
- We use the matching feature indices in M to select contextualized predictions and target local features from C^s and F^t, respectively
- Our objective is to **maximize agreement between contextualized and local embeddings** by minimizing the margin ranking loss defined as,

$$\mathcal{L} = \sum_{(i,j)\in M} \max\left(0, -\lambda\sigma\left(C_i^s, F_j^t\right) + \sigma\left(C_i^s, F_{\text{neg}}^t\right) + \mu\right), \tag{2}$$

where μ is the margin, $\sigma(a,b) = \frac{xy}{\|x\|_2\|y\|_2}$ is the cosine similarity and $\|\cdot\|_2$ is the ℓ_2 norm

- To create F_{neg}^t, (1) compute similarities between C^s and local representations from F^t, (2) select the top-k highest score representations from F^t, excluding the most similar one, and (3) average the resulting vectors.

Results

Obj. detection and segmentation on COCO (R50-FPN)

Method	ep	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mb}	AP^{mb}_{50}	AP^{mb}_{75}
Supervised	100	38.9	59.6	42.7	35.4	56.5	38.1
Rand init	–	32.8	51	35.3	28.5	46.8	30.4
DenseCL	200	39.4	59.9	42.7	35.6	56.7	38.2
ReSim	200	39.3	59.7	43.1	35.7	56.7	38.1
PixPro	400	39.8	59.5	43.7	36.1	56.5	38.9
SetSim	200	40.2	60.7	43.9	36.4	57.7	39
VICRegL	300	37.3	57.6	40.7	34.1	54.7	36.5
CLoVE	200	40.8	60.5	45.0	36.8	57.6	39.8
	400	**41.2**	**61.1**	**45**	**37.1**	**58.1**	**40.1**

Instance segmentation on Cityscapes (R50-FPN)

Method	ep	AP	AP_{50}
Supervised	100	26.5	52.9
Rand init	–	19.9	40.7
DenseCL	200	33.1	61.7
PixPro	400	35.8	63.7
VICRegL	300	29.8	58.5
SlotCon	200	35.2	63.8
CLoVE	200	35.7	64.1
	400	**37.2**	**65.3**

Keypoint detection on COCO (R50-FPN)

Method	ep	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
Supervised	100	65.3	87	71.3
Rand init	–	63	85.1	68.4
DenseCL	200	66.3	87.1	71.9
PixPro	400	66.6	87.2	73.0
ReSim	200	66.3	87.2	72.4
SetSim	200	66.7	**87.8**	72.4
SlotCon	200	66.5	87.5	72.5
CLoVE	200	66.9	87.5	73.2
	400	**67.0**	87.4	**73.3**

Acknowledgements

We thank Sigma2 (the National Infrastructure for High Performance Computing and Data Storage in Norway), Project NN8104K; the RCN Centre for Research-based Innovation funding scheme (grant no. 309439); and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.